行政院國家科學委員會專題研究計畫成果報告

使用機械學習技術來抽取文件自動分類系統中之分類規則之研究

A Study on Extracting Classification Rules in Automatic Document Classification Systems by Using Machine Learning Techniques

計畫編號：NSC 87-2213-E-032-016

執行期限：86 年 8 月 1 日至 87 年 7 月 31 日

主持人：洪文斌　淡江大學資訊工程系 (E-mail: horng@cs.tku.edu.tw)

一、中文摘要

自從 Maron 於 1961 年提出首篇的文件自動分類的論文以來，傳統的分類方法不外乎機率模式與向量模式。近年來的研究也加入了統計分析、專家系統、自然語言處理、和類神經網路等先進的技術，以提高分類的正確性。以上所提的諸方法中，其對文件自動分類而言，均可視為黑箱，因其分類行為或分類規則無從得知。本研究利用機械學習技術中之 Quinlan 的 C4.5 決策樹(decision trees)來抽取文件自動分類系統中之分類規則，期使文件自動分類系統之分類行為透明化，而人們可藉由所抽取之分類規則進一步來驗證文件自動分類之正確性。在本研究中，我們採用 *ACM Computing Reviews* 的分類法作為分類的依據。我們從該期刊共收錄了 56 個中類別，6424 篇論文為實驗用資料。再以其中的論文題目和出處當作該文件的素描。取其中十分之一為測試資料，其餘為訓練資料。我們從訓練資料中利用 Quinlan 的決策樹共抽取出 1162 條分類規則。再利用此分類規則分別對訓練文件及測試文件做分類，實驗結果分別為：訓練資料召回率為 67.7%，測試資料為 45.5%。若將上述規則再精簡成 290 條分類規則，則訓練資料召回率為 52.3%，測試資料為 43.0%。

關鍵詞：文件自動分類，資訊檢索，機械學習

Abstract

Since Maron proposed the first paper on automatic document classification in 1961, traditionally there are two approaches used: the probability model and the vector space model. Recent research also includes the advanced techniques of statistics, expert systems, natural languages processing, and artificial neural networks to enhance the correctness of document classification. However, all of the aforementioned methods could be regarded as black boxes for automatic document classification, because there are no ways to obtain their classification behaviors or classification rules. This project uses Quinlan's C4.5 decision trees of machine learning techniques to extract classification rules from automatic documents classification systems. In this research, the classification system of *ACM Computing Reviews* is based on. Totally 6424 papers, including 56 classes, are collected from it. The title and its source of each paper are used as its document profile. Among the collected papers, 10% of them are used as test data, and the remaining are used as training data. Totally, there are 1162 classification rules extracted from the training data using Quinlan's decision trees. These extracted classification rules are then used to categorize the training documents and test documents, respectively. The experiment results show that, the recall rates of training data and test data are 67.7% and 45.5%, respectively. If the above rules are

further simplified into 290 classification rules, the recall rates of training data and test data are 52.3% and 43.0%, respectively.

Keywords: automatic document classification, information retrieval, machine learning

二、緣由與目的

隨著資訊時代的來臨，網路的發達，資訊正以等比級數般的數量在激增。要在這龐大的資料中，找尋相關的資訊，確非易事。因此，文件自動分類的研究便應運而生。文件自動分類的目的即是利用電腦將性質相近的的資料或文件排放在一起，以提高文件分類的正確性與一致性，便於使用者能夠快速地檢索到相關的資訊。

Maron [10] 於 1961 年發表的論文，應該是文件自動分類領域中最早的文獻。Maron 認為，對於所要分類的文件，我們可以從文件中的某些詞找到分類的線索，稱之為關鍵詞(keywords)。若電腦也能從文件中自動找出這些關鍵詞，那麼便可以做到所謂的自動分類。在該論文中，他首先挑選了 405 篇文件，其中的 260 篇是訓練資料，另外的 145 篇是測試資料。每篇均取其摘要當作文件的素描。結果，在所有訓練資料中，共得到 3263 個不同的詞。其次，做關鍵詞篩選，把這些詞中去掉頻率最高的 55 個，及只出現一次或兩次的詞，便剩下 1088 個詞。再根據 Entropy 公式來計算，看這些詞在文件的分佈情形。只有分佈不平均的才有分類的價值，所以把平均者去掉。最後就只剩 90 個詞，也就是關鍵詞。他採用機率模式來作文件分類的實驗。結果顯示，在扣除不含關鍵詞及只含一個關鍵詞的文件後，訓練資料中有 84.6%的召回率，而測試資料中也達到 51.8%。(召回率即系統辨識正確之文件數和文件總數之比率。)

之後，陸續還有許多學者提出不同的作法，像Borko和Bernick [3] 延續了Maron的實驗，嘗試用向量模式來做分類；Kar 和 White [7] 的實驗，提出了第二選擇類別來提高正確率及循序的演算法來節省時間及空間；Kwok [8] 的分類實驗，除了論文題目及摘要外，他還利用論文所引用參考文獻的題目作為分類之用；以及 Hamill 和 Zamora [4] 的分類實驗(以下簡稱為 Hamill 方法)，提出了只用文件的題目來做分類。近年來的研究也加入了統計分析、專家系統、和自然語言處理等先進的技術，以提高分類的正確性 [2][5][6]。

類神經網路自從 1980 年代復甦以來，已吸引了許多研究者的投入，從事基礎性理論的研究與實務性的應用，並獲致相當的成功。近年來，也開始有人嘗試利用類神經網路來作文件的聚類(clustering)和分類 (classification) 之研究。MacLeod 和 Robertson [9] 利用非監督式學習網路中類似自適應共振理論網路 (Adaptive Resonance Theory Networks)模式來作文件之自動聚類研究。而 Yang [13] 依據傳統的相似性公式(cosine similarity measure)來設計一三層的類神經網路架構，其輸入層為文件中所出現的單字，輸出層為分類的類別，而隱藏層為訓練之文件。輸入層到隱藏層間的連結加權值(link weights)根據相似性公式適當的給定，而隱藏層到輸出層之間的連結加權值則為一條件機率 P(類別|文件)。對一測試文件相對於某一類別的相關性可輕易地由上述的相似性與條件機率相乘而得。最近，洪文斌與黃連進 [1] 也利用倒傳遞神經網路模式 [12] 來嘗試文件自動分類，並獲致不錯之成果。然而，上述諸方法中，無論是傳統的機率與向量模式或是先進的類神經網路模式，其對文件自動分類而言，均可視為是黑箱作業，因他們的分類行為或分類規則無從得知。

本研究計畫的主要目的，是嘗試利用機械學習技術來抽取文件自動分類系統中之符號分類規則，期使文件自動分類系統之分類行為透明化，而人們可藉由所抽取之分類規則進一步來驗證文件自動分類之正確性。在本研究中，我們採用了傳統的 Quinlan 的C4.5 [11] 決策樹(decision trees)來抽取文件分類的規則。在此計畫中，我

們也與傳統的機率模式與向量模式的分類結果，和倒傳遞類神經網路模式的分類結果作一比較。

三、實驗方法及步驟

在本文件自動分類研究中，我們採用 *ACM Computing Reviews* 的分類作為分類的依據。其分類系統共有 11 個大類和 80 個中類。我們從該期刊上，收錄了自西元 1986 年 1 月份起至 1997 年 6 月份止，共 67 個中類別，6507 篇論文。其中有 11 個中類別所含之論文數少於 10 篇，沒有足夠資訊以為訓練，將之刪除，並將有重覆出現之論文 49 篇去除，剩餘 56 個中類別，6424 篇論文為實驗用資料。再以其中的論文題目和出處當作該文件的素描。依論文收錄之順序，每第十篇取作為測試資料，計有 643 篇文件；其餘有 5781 篇為訓練資料。

在實驗前，首先將所有英文單字中，名詞單複數，及動詞單複數和其現在分詞、過去式、過去分詞取其原型；並將文件出處視為一英文單字。在關鍵詞的選取中，我們從訓練資料中，先去除 235 個 stop words (如 about, but, in, not, ...) 及只包含一個字元的單字，再以單字至少出現 5 次以上和其 Entropy 值小於等於 $\log_{10}(20)$ 為標準，共選出了 1146 個關鍵詞。在原先訓練文件中，有 26 篇沒有出現任何關鍵詞，去除後剩下 5755 篇；相同的，原先測試文件中，有 6 篇沒有出現任何關鍵詞，去之剩下 637 篇，此為以下實驗用之資料。

以下，我們對 C4.5 決策樹在抽取分類規則的原理做一簡單的介紹。Quinlan 的決策樹是傳統符號機械學習方法中，最具代表性與實用性。它從大量的輸入範例中，利用歸納推論建構而成。在建構決策樹的過程中，不斷地應用資訊理論中之 Entropy 公式來選取適當的屬性，再以其屬性值作為決策樹的分枝，將原先的輸入範例依其屬性值歸入相對的分枝內，如是重複地選取適當的屬性，再分枝，直至該分枝內所含的範例均屬於同一類別為止。當決策樹建好之後，我們可從樹根開始往各樹葉走，如是便可得出一組決策樹的分類規則，以所經過的屬性值為規則的條件，而該樹葉的類別為規則的結論。再經過一些簡化過程以去除不必要的條件，便可得到更精簡的分類規則。於是，決策樹的分類行為變成透明化，而我們可進一步去檢視所得的分類規則。我們再利用這些分類規則對原先的訓練範例和測試範例作分類，以得出其分類的正確性

在 Quinlan 的 C4.5 決策樹的文件自動分類實驗上，我們將所選出的 1146 個關鍵詞當成每個文件的屬性，若該關鍵詞有出現在文件中，則其屬性值為 Yes；若否，則為 No。我們利用訓練文件產生出 1162 條分類規則，再利用此分類規則分別對訓練文件和測試文件做分類，其召回率分別為：訓練文件 67.7%，測試資料 45.5%。

然而，經由以上所產生的分類規則，其條件部分，可能有數十個屬性以上所構成，過於複雜而難以瞭解。因此，再將以上規則進一步精簡為 290 條規則，取其中數條規則列舉於下，以供參考：

Rule 19:
music = Yes
→ class J.5 (Arts and Humanities)

Rule 26:
wafer = Yes
→ class B.7 (Integrated Circuits)

Rule 30:
curriculum = Yes
→ class K.3 (Computers and Education)

Rule 72:
R055 = Yes
vision = Yes
→ class I.2 (Artificial Intelligence)

Rule 84:
image = Yes
synthesis = Yes
→ class I.3 (Computer Graphics)

Rule 204:
extract = Yes
information = Yes
→ class H.3 (Information Storage and

Retrieval)

Rule 269:

debugger = Yes

→ class D.2 (Software Engineering)

其中，第 72 條規則中之 R055 為文件之期刊出處：*Computer Vision, Graphics, and Image Processing*。其義為：若文件出處為上述期刊並且在文件題目中出現 vision，則該文件應分類為 I.2，即人工智慧類別 (Artificial Intelligence)。若以精簡的 290 條分類規則來分類，則訓練資料召回率為 52.3%，測試資料為 43.0%。

此外，我們並實驗了傳統的機率模式、向量模式、Hamill 的分類方法、以及倒傳遞神經網路以為比較。我們將以上實驗結果表列如下：

表一：各種文件分類實驗方法的召回率

文件分類實驗方法	訓練資料	測試資料
Quinlan's C4.5 (原始)	67.7%	45.5%
Quinlan's C4.5 (精簡)	52.3%	43.0%
倒傳遞網路	70.7%	57.1%
機率模式	82.7%	39.4%
向量模式	55.4%	43.0%
Hamill 方法	63.5%	55.3%

四、結果與討論

在前一節的實驗數據顯示，Quinlan 的 C4.5 實驗方法中，在測試資料方面雖不及倒傳遞網路及 Hamill 方法，然而，它卻可用 290 條已精簡過的分類規則達到 43.0% 的召回率，此分類規則已將文件分類的方式完全透明的表達出來，便於人類專家閱讀與驗證其正確性。在本實驗中，各種方法的正確性並不很高，此可能為實驗用文件素描只包含論文題目及其出處，平均每一篇文件只包含了 4.8 個關鍵詞，由於資訊相當有限，故在各種實驗中，測試資料召回率無法超越 60%。

五、參考文獻

[1] 洪文斌和黃連進，"使用類神經網路來作文件自動分類之研究"，*1998 分散式系統技術及應用研討會*，台南，民國八十七年五月，209－216 頁。

[2] M.J. Blosseville, M.J. Hebrail, M.G. Monteil, and N. Penot, "Automatic Document Classification: Natural Language Processing, Statical Analysis, and Expert System Techniques Used Together," in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, June 21-24, 1992, pp. 51－58.

[3] H. Borko and M. Bernick, "Automatic Document Classification," *Journal of the ACM*, Vol. 10, No. 1, 1963, pp. 151－162.

[4] K.A. Hamill and A. Zamora, "The Use of Titles for Automatic Document Classification," *Journal of the American Society for Information Science*, Vol. 31, November 1980, pp. 396－402.

[5] P.S. Jacobs, "Joining Statistics with NLP for Text Categorization," in *Proceedings of the 3rd Conference on Applied Natural Language Processing*}, Trento, Italy, March 31-April 3, 1992, pp. 178－185.

[6] P.S. Jacobs, "Using Statistical Methods to Improve Knowledge-Based News Categorization," *IEEE Expert*, Vol. 8, No. 2, April 1993, pp. 13－23.

[7] G. Kar and L.J. White, "A Distance Measure for Automatic Document Classification by Sequential Analysis," *Information Processing and Management*, Vol. 14, 1978, pp. 57－69.

[8] K.L. Kwok, "The Use of Title and Cited Titles as Document Representation for Automatic Classification," *Information and Management*, Vol. 11, 1975, pp. 201－206.

[9] K.J. MacLeod and W. Robertson, "A Neural Algorithm for Document Clustering," *Information Processing and Management*, Vol. 27, No. 4, 1991, pp. 337－346.

[10] M.E. Maron, "Automatic Indexing: An Experimental Inquiry," *Journal of the ACM*, Vol. 8, 1961, pp. 404-－417.

[11] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

[12] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning Internal Representation by Error Propagation," in *Parallel Distributed Processing*, Vol. 1, pp. 318－362.

[13] Y. Yang, "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ireland, 1994, pp. 13－22.